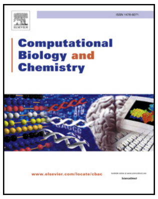
Computational Biology and Chemistry 80 (2019) 90–101

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac



Research Article

Predicting drug-target interaction network using deep learning model

Jiaying You[a,b,c], Robert D. McLeod[b], Pingzhao Hu[a,b,c,d,*]

[a] Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, Manitoba, Canada
[b] Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Manitoba, Canada
[c] George & Fay Yee Centre for Healthcare Innovation, University of Manitoba, Winnipeg, Manitoba, Canada
[d] Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, Canada

ARTICLE INFO

Keywords:
Deep learning
Drug repurposing
Feature integration
Drug-target interaction
LASSO models

ABSTRACT

Background: Traditional methods for drug discovery are time-consuming and expensive, so efforts are being made to repurpose existing drugs. To find new ways for drug repurposing, many computational approaches have been proposed to predict drug-target interactions (DTIs). However, due to the high-dimensional nature of the data sets extracted from drugs and targets, traditional machine learning approaches, such as logistic regression analysis, cannot analyze these data sets efficiently. To overcome this issue, we propose LASSO (Least absolute shrinkage and selection operator)-based regularized linear classification models and a LASSO-DNN (Deep Neural Network) model based on LASSO feature selection to predict DTIs. These methods are demonstrated for repurposing drugs for breast cancer treatment.
Methods: We collected drug descriptors, protein sequence data from Drugbank and protein domain information from NCBI. Validated DTIs were downloaded from Drugbank. A new similarity-based approach was developed to build the negative DTIs. We proposed multiple LASSO models to integrate different combinations of feature sets to explore the prediction power and predict DTIs. Furthermore, building on the features extracted from the LASSO models with the best performance, we also introduced a LASSO-DNN model to predict DTIs. The performance of our newly proposed DNN model (LASSO-DNN) was compared with the LASSO, standard logistic (SLG) regression, support vector machine (SVM), and standard DNN models.
Results: Experimental results showed that the LASSO-DNN over performed the SLG, LASSO, SVM and standard DNN models. In particular, the LASSO models with protein tripeptide composition (TC) features and domain features were superior to those that contained other protein information, which may imply that TC and domain information could be better representations of proteins. Furthermore, we showed that the top ranked DTIs predicted using the LASSO-DNN model can potentially be used for repurposing existing drugs for breast cancer based on risk gene information.
Conclusions: In summary, we demonstrated that the efficient representations of drug and target features are key for building learning models for predicting DTIs. The disease-associated risk genes identified from large-scale genomic studies are the potential drug targets, which can be used for drug repurposing.

1. Background

Drug development is usually a high cost and time-consuming procedure. It has been estimated that bringing a new drug into market successfully could take 10 years with a cost on the order of $Billion (Health et al., 2014). Novel drug discovery technologies, such as structure-based drug design, combinatorial chemistry, high-throughput screening and genomics (Ashburn and Thor, 2004), have been applied to minimize the risk versus return trade-off in drug research and development, but the results are not so overly encouraging with only few successful products(Horrobin, 2001). Biopharmaceutical companies have been making efforts to increase the productivity by introducing

Abbreviations: ACC, amino acid composition; AUC, area under the curve; DC, dipeptide composition; DL, deep learning; DNN, deep neural network; DTI, drug-target interaction; FNR, false negative rate; FPR, false positive rate; GBM, gradient boosted machine with trees; GWAS, genome-wide association study; PRC, precision-recall curve; RBM, restricted Boltzmann machine; ReLU, rectified linear units; RF, random forest; ROC, receiver operating characteristic; SLG, standard logistic; SNP, single nucleotide polymorphism; SVM, support vector machine; TC, tripeptide composition; TNR, true negative rate; TPR, true positive rate

[*] Corresponding author at: Department of Biochemistry and Medical Genetics, Room 308 - Basic Medical Sciences Building, 745 Bannatyne Avenue, University of Manitoba, Winnipeg, Manitoba, R3E 0J9, Canada.

E-mail address: pingzhao.hu@umanitoba.ca (P. Hu).

https://doi.org/10.1016/j.compbiolchem.2019.03.016
Received 7 March 2019; Accepted 23 March 2019
Available online 25 March 2019

other new drug discovery methods. One example is to discover new uses of existing drugs, which is known as drug repositioning or drug repurposing (Gilbert et al., 2003).

Advantages in drug repurposing compared to traditional methods are significant in time and cost since the repurposed candidates have already been through clinical usage trials and thus could be substantially bypassed in terms of their safety concerns. These attractive characteristics lead to increasing interests for biopharmaceutical companies to scan the existing drugs for repurposing usages. It has been estimated that approximately 30% of FDA-approved new drug products were made available through drug repurposing.

The technologies for drug repurposing are developing rapidly. As the biological and experimental methods for drug repurposing are time-consuming, expensive and difficult, the shortcomings of traditional methods have led to increased interests in applying sophisticated computational methods, such as machine learning, to analyze the vast amount of available complex data for drug repurposing. Machine learning has been used in discovering patterns from diverse data sources and have demonstrated excellent versatility in diverse applications for many years, and these technologies have proved to be highly useful in identifying drug-target interactions (DTIs) (Ezzat et al., 2017). Many researches have applied machine learning technologies in drug repurposing. A study by Zhao et.al applied several machine learning approaches, such as support vector machine (SVM), random forest (RF), gradient boosted machine with trees (GBM) and logistic regression approaches as baseline models. These were compared to deep neural network models for drug repurposing in psychiatry using drug expression profiles, which express the transcriptomic changes when specific cell lines are affected with drugs or chemicals (Zhao and So, 2017). Given the interactions between drugs and genes, supervised learning has been applied to discover unknown genes that could be affected by compounds. Drugs that had high prediction probability values in the model were believed to be good candidates for repurposing.

Yamanishi et al. developed a kernel regression-based approach for predicting DTIs that can identify previously unknown DTIs from various types of biological data (Yamanishi et al., 2014), such as chemical structures, drug side effects, amino acid sequences and protein domains. They used these features to generate a kernel similarity matrix, where each descriptor corresponds to a drug-drug similarity or protein-protein similarity. The model assumed that similar compounds were likely to interact with similar proteins to generate their prediction model. Generally speaking, drugs with high similarities are considered to have shared target proteins and new interactions observed are used for drug repurposing. They applied their analyses to five diseases: Huntington disease, two subtypes of lung cancer, alcohol dependence and polysubstance abuse, and reported the most significant candidates repurposed for each disease.

Other DTI prediction methods, such as network-based and phenotype-based algorithms, are rapidly emerging. For example, Cheng et al. designed three network-based inference algorithms for DTI prediction (Cheng et al., 2012). Two of them were associated with drug similarity and protein similarity networks respectively, and the last one was simply based on the known DTIs. Generally speaking, for the first two approaches, the authors took the chemical compound structure and genomic sequence data to quantify the similarity between them and assumed that similar drugs may share the same targets and vise-versa. Although the results from the drug similarity network and protein similarity network seemed poor, the simplest network approach using the known DTIs showed the best result among the three methods.

Due to the high-dimensional and noisy data used in the machine learning approaches for drug repurposing (Napolitano et al., 2013), recently developed deep learning algorithms that were shown to have greater power than traditional machine learning algorithms for drug repurposing (Min et al., 2017; Leung et al., 2016). Wen et al. proposed a deep learning-based model DeepDTIs for predicting DTIs (Wen et al., 2017). The study first generated features by unsupervised pre-training, then applied a supervised deep learning algorithm to the known DTIs. By measuring the accuracy, area under the receiver operator characteristic (AUC), true positive rate (TPR, aka sensitivity) and false positive rate (TNR, aka specificity), the authors showed that the DeepDTIs model had the best performance compared with other baseline models including Bernoulli Naïve Bayesian (BNB) methods, Decision Trees (DT), and RFs. They also argued that the predicted DTIs with top probability scores can be used for repurposing drugs for some diseases. For example, 5-hydroxytryptamine receptor 1A, 2A, 1B, 2C, 1D were shown in drug Ziprasidone, which were proved to have high sequence homology with 5-hydroxytryptamine receptor 1C.

It is believed that drugs approved for diseases through genetic studies are twice as likely to be clinically accepted than other drugs (Pritchard et al., 2017). Since recent genome-wide associate studies (GWAS) have identified many potential causal genes for different complex diseases (MacArthur et al., 2017), there is interests to investigate the use of the disease risk genes as targets for drug repurposing. For example, Sanseau et al. constructed a list of 991 GWAS genes associated with different diseases and traits from the catalog of the published GWAS data. They found that 92 of these individual genes were newly identified as targets of drugs with known targets (Sanseau et al., 2012).

This study is motivated by previous findings and aims to develop a new deep neural network model for predicting DTIs. The model will integrate both drug structure and protein sequence information. Furthermore, we explore repurposing candidate drugs for breast cancer using the predicted DTIs with the breast cancer risk genes identified from large-scale genome-wide genetic studies.

2. Materials and methods

2.1. Datasets

2.1.1. Data collection

The drug and target sets were downloaded from Drugbank (Wishart et al., 2018). Drugbank is a bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information. Drugbank contains approved small molecule drugs,

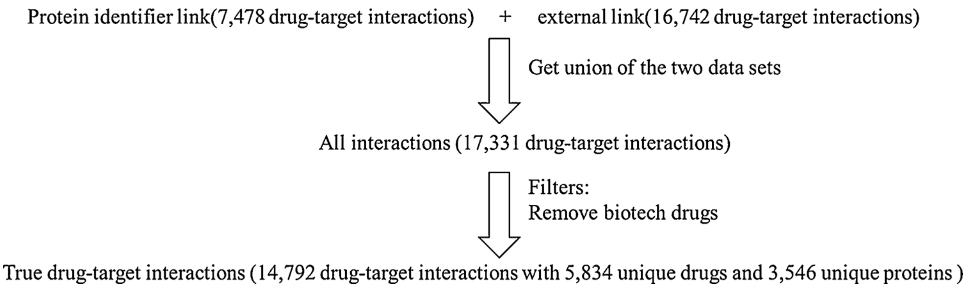

Protein identifier link(7,478 drug-target interactions)
+
external link(16,742 drug-target interactions)
Get union of the two data sets
All interactions (17,331 drug-target interactions)
Filters:
Remove biotech drugs
True drug-target interactions (14,792 drug-target interactions with 5,834 unique drugs and 3,546 unique proteins)

Fig. 1. Flowchart of true (known) drug-target interaction data collection. 14,792 known drug-target interactions were generated from Drugbank with 5834 unique drugs and 3546 unique proteins.

approved biotech drugs, nutraceutical and experimental drugs with non-redundant protein sequences linked to all the drug entries. As shown in Fig. 1, we used DTIs data under protein identifier (Version 5.0.10) and external target drug-uniprot link (Version 5.0.10). Biotech drugs were excluded since the Rcpi R package used for drug feature extraction is only applied to small molecular drugs. Thus, 14,792 known DTIs from Drugbank were extracted with 5834 unique drugs and 3546 unique proteins.

2.1.2. Drug and target feature representation

Chemical properties, topological properties, and geometrical properties are often referred as drug descriptors. All these can be considered as drug features. We downloaded drug structure information from Drugbank (https://www.drugbank.ca/releases/latest#structures). It should be noted that some drug structures are not available, thus only 5585 drugs were kept in this procedure. To get drug descriptors from the drug structure information, an online sever named online chemical database with modeling environment was used (Sushko et al., 2011). Drugs with unavailable descriptors from this sever were removed. We retrieved 2216 drug descriptors for each of the remaining 5134 drugs. To generate target/protein feature representations, we assumed that the complete information of a target protein is encoded in its sequence (He et al., 2010) and domains, thus we extracted protein features from the commonly recognized features-sequence descriptors and protein domains. The sequence and corresponding domain information of target proteins in our study were downloaded from Drugbank and NBCI (Wheeler et al., 2007) (National Center for Biotechnology Information Search database, https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) respectively. Protein sequence descriptors contain amino acid composition (ACC), dipeptide composition (DC) and tripeptide composition (TC). AAC is the statistic frequency of each amino acid. DC is the statistic frequency of every two amino acid combination while TC is the statistic frequency of every three amino acid combination. Domain information is represented as an adjacency matrix where 1 indicates an interaction and 0 indicates no interaction pairs of proteins and domains. In total, we extracted 11,943 protein features for each protein target. Therefore, each pair of drug-target/protein (D, P) can be represented by a feature vector as [D_1, ...,D_{2216}, P_{ACC1}, ...,P_{ACC20}, P_{DC1},...P_{DC400}, P_{TC1}, ...P_{TC8000}, $P_{domain1,...,}P_{domain3523}$], which contains 2216 drug features: D_1, ...,D_{2216}, 11,943 target features with 20 ACC expressions P_{ACC1}, ...,P_{ACC20}, 400 DC expressions P_{DC1},...P_{DC400}, 8000 TC expressions P_{TC1}, ...P_{TC8000}, and 3523 domains $P_{domain1,...,}P_{domain3523}$.

After all drug and target feature information was generated, 13,168 of the 14,792 known DTIs with 5132 unique drugs and 3184 unique proteins were collected in our study as shown in Fig. 2. All these 13,168 known DTIs were treated as positive samples. Since the number of all possible DTIs from the collected drug and target lists is 18,138,422 (5134*3533) (Fig. 2), 18,125,254 of which (do not include the positive DTIs set) were considered as unknown DTIs. These can be potential negative or positive DTIs.

2.1.3. Construction of negative DTIs

Although the large number of unknown interaction pairs are more likely to be negative since the number of no (negative) interaction pairs is far more than the number of interaction pairs(Wen et al., 2017), it is possible that some of these pairs interact. Hence, instead of randomly selecting negative samples from the remaining 18,125,254 unknown pairs, we proposed a novel method to select the most likely negative DTIs. The assumption of this method is based on “guilt-by-association”, which indicates similar drugs may share similar targets and vice versa (Luo et al., 2017). Based on this intuition, we calculated the drug similarity measured by Tanimoto coefficient for each pair of the drugs using R package Rcpi and drugs with Tanimoto coefficients > 60% were inferred as similar in this study (Bajusz et al., 2015). We used a well-known method to map molecular structures with bit strings (i.e. molecular fingerprints). The Tanimoto coefficients between drugs for continuous variables as shown in Eq. (1) and dichotomous variables as shown in Eq. (2) were respectively defined by Bajuse et al.(Bajusz et al., 2015)

$$S_{A,B} = \frac{\left|\sum_{j=1}^{n} x_{jA}x_{jB}\right|}{\left|\sum_{j=1}^{n} (x_{jA})^2 + \sum_{j=1}^{n} (x_{jB})^2 - \sum_{j=1}^{n} x_{jA}x_{jB}\right|} \quad (1)$$

$$S_{A,B} = \frac{c}{a + b - c} \quad (2)$$

Here, S donates similarity between drug A and drug B, x_{jA} donates j-th feature of A and a is the number of on bits in molecule A. x_{jB} donates j-th feature of B and b is the number of on bits in molecule B, while c is the number of bits that are on in both molecules. n is the number of features.

It is believed that similar drugs are more likely to be functionally correlated, thus, targets that interact with one drug are also considered to interact with its similar drugs, and these DTIs were referred as potential interaction pairs in our study as shown in Fig. 3. Protein similarities were also calculated using Protr R package (Xiao et al., 2015). They were derived by protein sequence alignment, which is a way to identify regions of similarity of proteins based on their functional, structural, or evolutionary relationships. After obtaining the similarities between proteins, the same filtering rule was applied to find potential

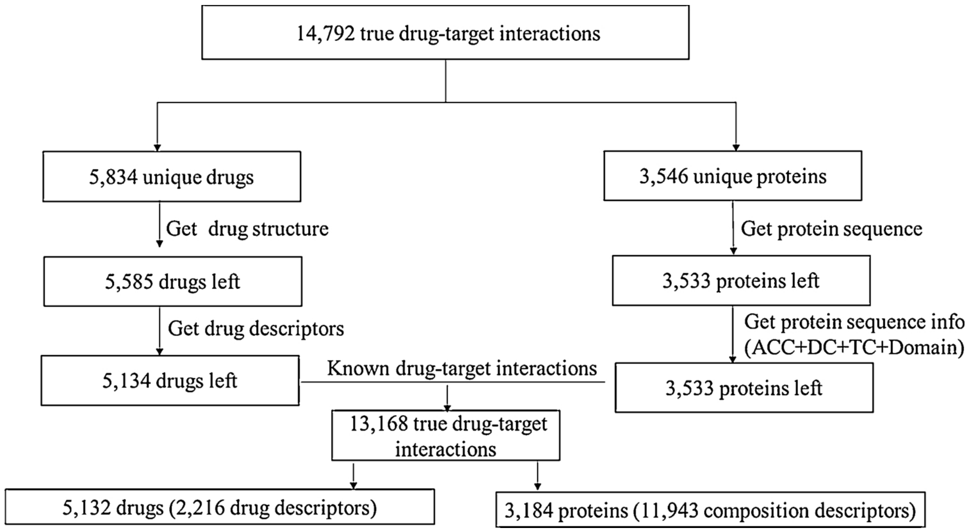
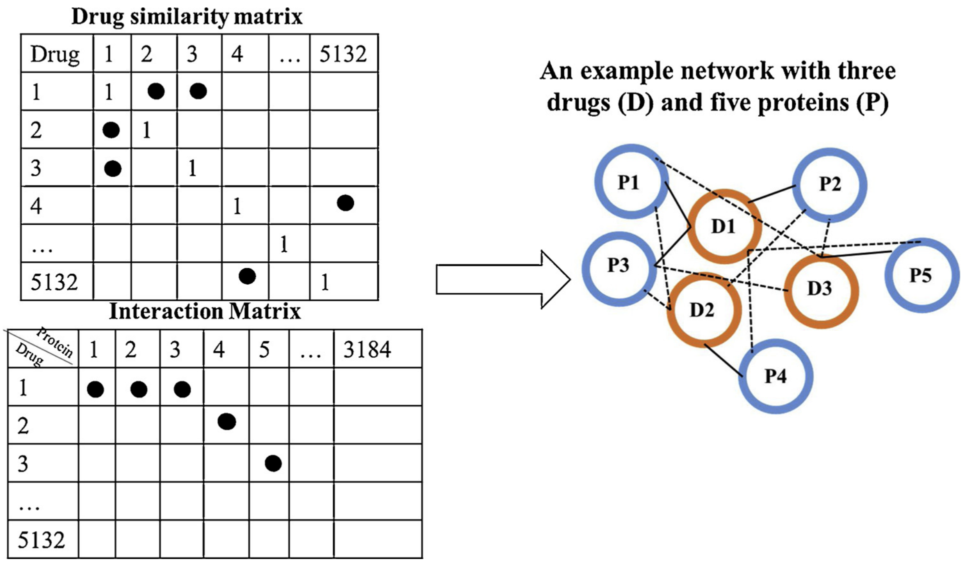

14,792 true drug-target interactions
5,834 unique drugs
3,546 unique proteins
Get drug structure
Get protein sequence
5,585 drugs left
3,533 proteins left
Get drug descriptors
Get protein sequence info
(ACC+DC+TC+Domain)
5,134 drugs left
Known drug-target interactions
3,533 proteins left
13,168 true drug-target interactions
5,132 drugs (2,216 drug descriptors)
3,184 proteins (11,943 composition descriptors)

Fig. 2. Flowchart of feature generation. 13,168 known drug-target interactions with 5132 unique drugs and 3184 unique proteins were filtered out as our positive DTIs with both drug and protein features. 18,125,254 drug-target pairs were considered as unknown interactions. These can be potential negative or positive DTIs we will predict.


Drug similarity matrix
Interaction Matrix
An example network with three drugs (D) and five proteins (P)
P1
P2
P3
P4
P5
D1
D2
D3

Drug	1	2	3	4	…	5132
1	1	●	●			
2	●	1				
3	●		1			
4				1		●
…					1	
5132				●		1

Protein Drug	1	2	3	4	5	…	3184
1	●	●	●				
2				●			
3					●		
…							
5132							

Fig. 3. Procedure of finding potentially interacted drug-target pairs by drug similarities. Bold dots in the similarity matrix indicated the Tanimoto coefficients of the drug pairs larger than 0.6. Thus, the corresponding drugs (D) were referred as being similar, such as D_1 and D_2, D_1 and D_3. The dots in the drug-target interaction matrix presented the known drug-target interactions. Based on the known interactions, the example network was generated (the solid edges), where each orange node was a drug and each blue node was a protein. Based on the known interactions and the drug similarity matrix, we inferred the potential true drug-target interactions (the dash lines) in the network. We can infer the potential true drug-target interactions based on the protein similarity matrix using the same procedure. Hence, all other non-interacted drug-target pairs can be treated as potential negative drug-target interactions.

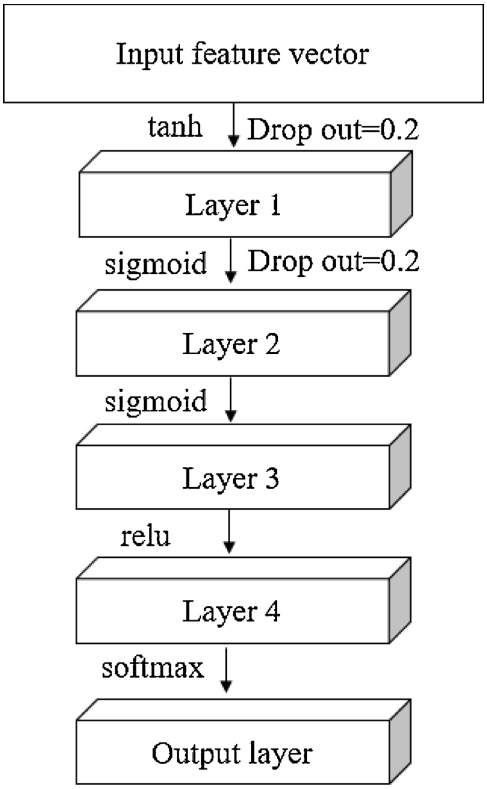


Fig. 4. Standard DNN structure. Positive and negative DTIs with all features were fed into the standard deep neural network model with four hidden layers. The output layer outputted the probability of a drug-target pair to be a true DTI.

DTIs based on the protein similarity scores. That is, if the similarity score between two proteins (sequence identity score) is above 40% (Luo et al., 2017), it will be assumed that these proteins share the same drug interactions. We followed the similar procedure to identify potential DTIs by protein similarity. After we removed all potential positive DTIs, we randomly selected 13,168 negative DTIs (samples) from the remaining unknown DTIs.

2.2. Classification models

2.2.1. LASSO-based classification model

Using the constructed positive and negative DTIs to predict new DTIs, we first implemented standard logistic (SLG) and LASSO (Least absolute shrinkage and selection operator)-based regularized linear classification (Friedman et al., 2010) models. SLG is a machine learning algorithm used for binary classification. The aim is to find the optimized hyperplane (Eq. (3)) that correctly classifies the positive and negative DTIs.

$$z = \theta^T x \tag{3}$$

where x donates feature space and the optimized parameter vector? ? fits a curve to separate the training data. To simplify the measurement of its output, we considered the sigmoid function (Eq. (4)) to represent the probability of a DTI (y) given one sample x:

$$p(y = 1|z) = g(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

Overfitting might be noticeable in the model analysis since we had more than 14,000 features representing drugs and proteins while there were a relatively small number of samples (positive and negative DTIs). Thus it is necessary to select informative features for building the SLG-based classification model to alleviate the overfitting problem. Since the majority of our features were category variables, feature selection based on the training data set was implemented using Wilcoxon rank sum test. The test is a nonparametric statistical test that could be applied when the data population cannot be assumed to follow a normal distribution. It is a simple test ranking all samples (DTIs) from two classes (positive and negative DTIs) and obtaining sum ranks of each class. The significance of each feature can be measured by its P-value, which is the probability of obtaining the estimated parameter θ value under the null hypothesis of θ = 0. Thus, to decrease the effect of those less significant features (measured by P-value mentioned above), the LASSO model was also implemented. This reduces the effect of features, which are referred as "not significant" as estimated by their P-values. In other words, only features considered "significant" could have contributions to the optimized hyperplane. The decision boundary is now added a penalty term to realize the assumption that the coefficients of those not crucial features are set to zero as shown in Eq. (5).

$$z = \theta^T x + \lambda \sum |\theta| \tag{5}$$

Here, λ is selected to minimize the output of sample errors by a gird search using cross-validation technology.

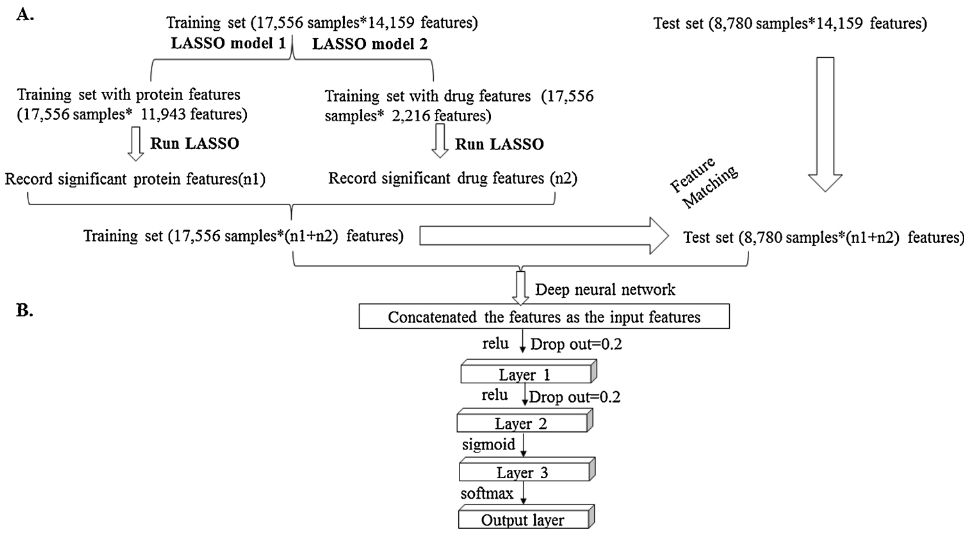

A.
Training set (17,556 samples*14,159 features)
LASSO model 1
LASSO model 2
Test set (8,780 samples*14,159 features)
Training set with protein features (17,556 samples* 11,943 features)
Training set with drug features (17,556 samples* 2,216 features)
Run LASSO
Run LASSO
Record significant protein features(n1)
Record significant drug features (n2)
Feature Matching
Training set (17,556 samples*(n1+n2) features)
Test set (8,780 samples*(n1+n2) features)
B.
Deep neural network
Concatenated the features as the input features
relu
Drop out=0.2
Layer 1
relu
Drop out=0.2
Layer 2
sigmoid
Layer 3
softmax
Output layer

Fig. 5. The architecture of the LASSO-DNN network. The new DNN structure includes two components. **A: Feature extraction.** Two LASSO models were applied to protein features and drug features respectively. n1 protein features and n2 drugs features were extracted through the LASSO models. **B: Deep neural network**. Data with concatenated n1 + n2 features were generated before feeding into the deep neural network model with three hidden layers. The output layer outputted the probability of a drug-target pair to be a true DTI.

Table 1
Overall accuracies and AUCs of the SLG models using the test set.

SLG model	Numbers of top features selected by Wilcoxon rank sum test				
	100	200	500	1000	2000
Accuracy	0.64	0.66	0.70	0.72	0.75
AUC	0.70	0.73	0.77	0.80	0.81

In our study, we applied SLG models with the top 100, 200, 500, 1000, 2000 of all 14,159 features selected by Wilcoxon rank sum test respectively. We also tried to use more than 2000 features in the SLG models, but the models could not be fitted due to the large number of features and small number of samples. We implemented five LASSO models. Feature vectors for each of the five LASSO models (M1, M2, M3, M4 and M5) were represented as:

$M1$: $[D_1,\ \ldots, D_{2216},\ P_{ACC1},\ \ldots, P_{ACC20},\ P_{DC1},\ldots P_{DC400},\ P_{TC1},\ \ldots P_{TC8000},\ P_{domain1,\ldots,}P_{domain3523}]$

Table 2
Accuracies and AUCs of the LASSO models for both the training and the test sets in the five models.

		M1	M2	M3	M4	M5
10-fold cross-validation of training set	**Accuracy**	0.75	0.63	0.69	0.74	0.74
	AUC	0.82	0.69	0.76	0.81	0.82
Test set	**Accuracy**	0.75	0.64	0.70	0.75	0.75
	AUC	0.83	0.69	0.77	0.83	0.83

$M2$: $[D_1,\ \ldots, D_{2216},\ P_{ACC1},\ \ldots, P_{ACC20}]$

$M3$: $[D_1,\ \ldots, D_{2216},\ P_{DC1},\ldots P_{DC400}]$

$M4$: $[D_1,\ \ldots, D_{2216},\ P_{TC1},\ \ldots P_{TC8000}]$

$M5$: $[D_1,\ \ldots, D_{2216},\ P_{domain1,\ldots,}P_{domain3523}]$

We built these models using R packages e1071(Meyer et al., 2014) for SLG and glmne (Friedman et al., 2010) for LASSO based on the training set. For the LASSO models, we performed 10-fold cross-validation of the training set to select the optimized parameter λ.

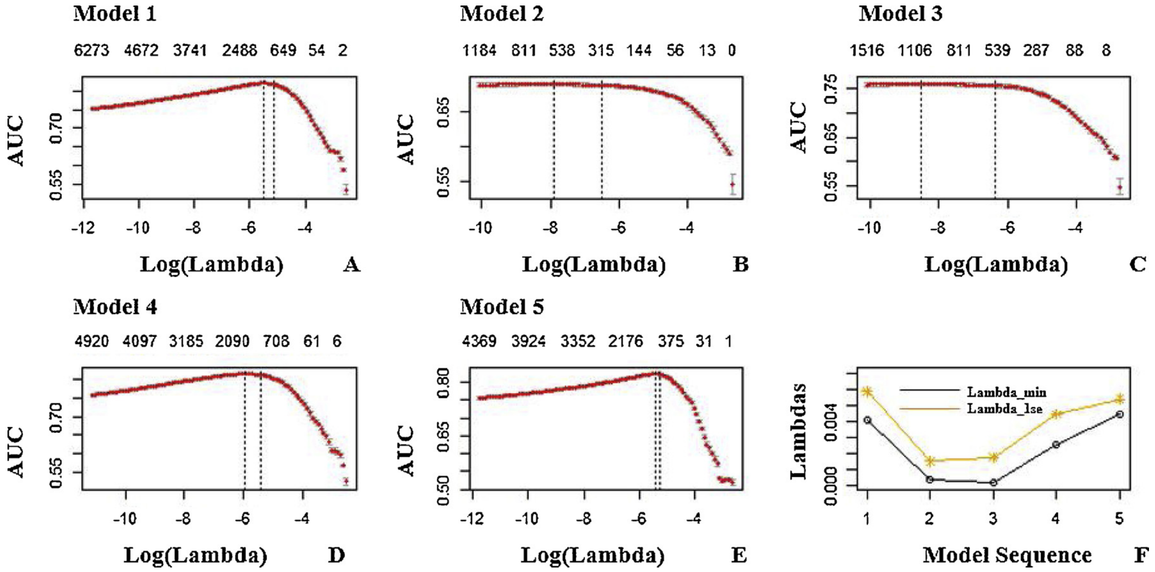


Fig. 6. Lambda selection in the five LASSO models using the training set. A–E showed AUCs of the models on training data using various Lambdas. In this study, the largest Lambda that produced the AUC within 1 standard error of the maximum AUC (Lambda_1se) was selected to avoid overfitting problem. F showed the values of Lambda_1se and Lambda_min for the five models.

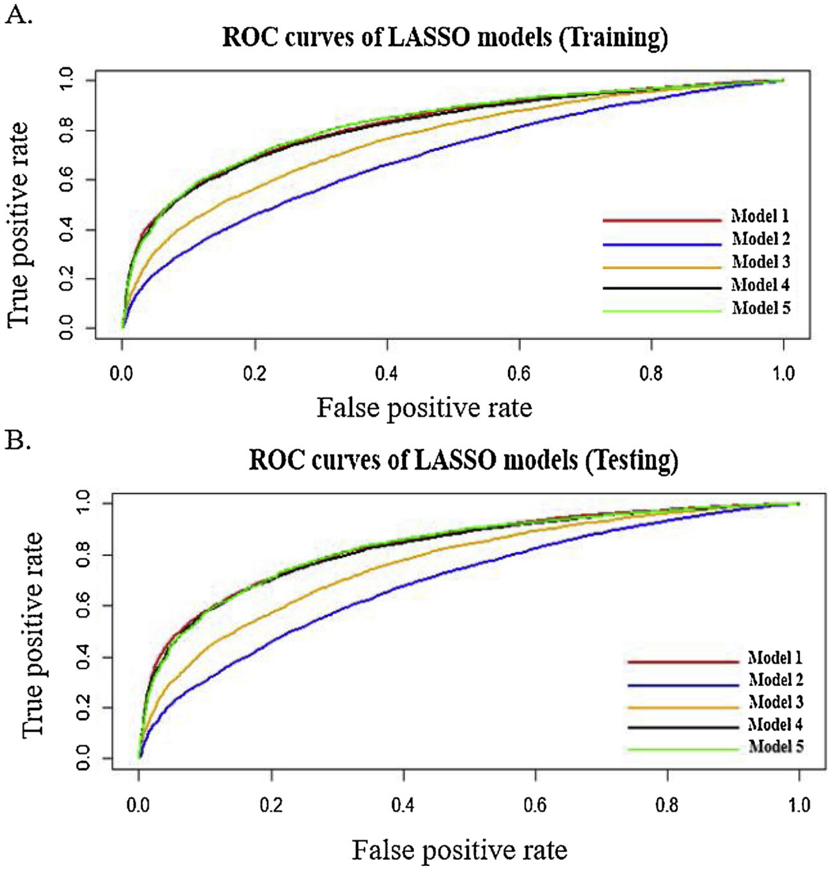


Fig. 7. ROC curves of the LASSO models for the training and test sets of the five models. A: the ROC curves of the five models in the training set; B: the ROC curves of the five models in the test set.

Table 3
Accuracies and AUCs of the SVM models using various values of parameter C.

	SVM Models				
C Values	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
Accuracy	0.75	0.77	0.77	0.77	0.77
AUC	0.82	0.84	0.84	0.83	0.83

Table 4
Accuracies, AUCs and the number of significant features of the two LASSO models using the training set.

Performance on training set using LASSO		Model 6	Model 7
Lambda_min	**Accuracy**	0.716	0.623
	AUC	0.754	0.659
	Number of significant features	1078	883
Lambda_1se	**Accuracy**	0.714	0.620
	AUC	0.752	0.657
	Number of significant features	867	596

2.2.2. Standard DNN classification model

A deep neural network is a kind of artificial neural network (ANN) with more hidden layers and more complex computations between input and output layers, and thus are designed to recognize more complicated patterns. The basic units of deep neural network are neurons, which are supposed to find patterns just like in human brains do. Deep neural networks have several associated optimization algorithms, such as Adaptive Moment Estimation (Adam)(Kingma and Ba, 2015) and techniques, such as dropout(Srivastava et al., 2014). These have been commonly used in applications, such as image classification and compound-protein interaction prediction (Hamanaka et al., 2017).

In this study, instead of using traditional gradient descent for fitting the model, we chose to use a modified Adaptive Moment Estimation (Adam) optimizer. The Adam algorithm is basically a combination of momentum and Root mean square prop (RMSprop) optimization methods. The momentum term is widely used in various neural network architectures to improve model performance. The momentum term allows the current direction to be changed somewhat based on the previous directions but not totally determined by gradient descent, which may help the model from getting stuck in a local minima. The percentage of current direction influenced by previous paths and gradient descent will be implemented through a new parameter: the momentum term. Adding the momentum term allows us to change the proportion of decrease coming from a descent calculation, which can be controlled by the performance on the training set. Meanwhile RMSprop optimization deceases the chance of oscillations in training performance and helps to learn the patterns faster.

In our standard DNN model, we included 6 layers: one input layer, 4 hidden layers and one output layer. We used various activation functions such as tanh, sigmoid, relu and SoftMax. The standard DNN model architecture with dropout used in our study is presented in Fig. 4.

2.2.3. LASSO-DNN classification model

2.2.3.1. Feature extraction for LASSO -DNN classification model. To improve the performance of the standard DNN model, we combined the LASSO and the DNN models (LASSO-DNN model) together. To do this, we first extracted drug- and protein-specific features from the training data, which had associated with it the DTI status using the LASSO models (Fig. 5A). We built two LASSO models on training data using R packages glmnet (Friedman et al., 2010) based on the drug feature set and protein feature set respectively. The features of these two models were as follows:

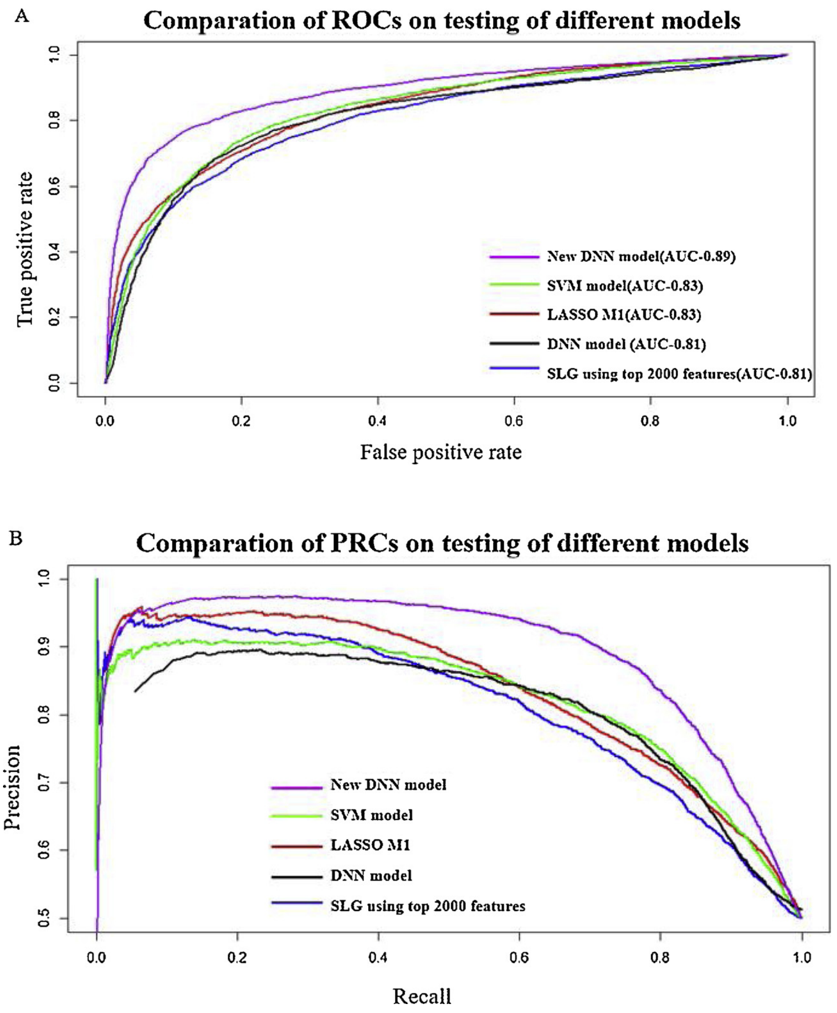


Fig. 8. Model performance comparisons. A: ROC Curves. ROC curves of the test set using the SLG, LASSO M1, SVM, standard DNN, and new DNN (Lasso-DNN) models. **B: PRC curves.** PRC curves of test set using the SLG, LASSO M1, SVM, standard DNN, and new DNN (Lasso-DNN) models.

Table 5
Accuracies and AUCs of SLG, LASSO M1, SVM, standard DNN and LASSO-DNN models.

Performance Comparisons	Accuracy	AUC
SLG	0.75	0.81
LASSO M1	0.75	0.83
SVM	0.77	0.83
Standard DNN	0.76	0.81
New DNN (LASSO - DNN)	0.81	0.89

$M6$: $[D_1, \ ..., D_{2216}]$

$M7$: $[P_{ACC1}, ..., P_{ACC20}, P_{DC1}, ...P_{DC400}, P_{TC1}, ...P_{TC8000}, P_{domain1,...,}P_{domain3523}]$

We performed 10-fold cross-validation to explore the optimization fit on the training set, and the features that showed statistical significance were recorded for the deep neural network implementation as illustrated in Fig. 5**A**.

2.2.3.2. Model design. The LASSO-DNN model architecture used in our study is presented in Fig. 5**B**. We kept features that showed significance in the LASSO model's training set and matched these features in the test set. We used relu, sigmoid and softmax as activation functions in different layers of the network and applied a dropout of 20% randomly from the input layer to the first hidden layer, and from the first hidden layer to the second hidden layer. Dropout was used to avoiding overfitting.

2.2.4. Software and model parameters

In the LASSO-DNN model implementation, we used TensorFlow library (Abadi et al., 2016). We concatenated significant features extracted from the two LASSO models fitted for the drug-specific and protein-specific features respectively. We trained our network using learning rate = 0.00001, with a dropout rate = 0.2 in the first two hidden layers and L2 regularization.

2.2.5. SVM classification model

As a comparison with our new LASSO-DNN model, a SVM was also built to classify the positive and negative DTIs. The SVM was based on the Python package sklearn. We used a linear kernel and the cost parameter C was set to 0.5 after tuning a range of different cutoffs of C.

2.2.6. Model selection strategy

We split our samples (positive and negative DTIs) into 2/3 and 1/3 of all samples as our training and test datasets respectively. For each of the training and test sets, we ensured there were the same numbers of positive and negative DTIs. The training set was used for model selection and the test set was used for model performance evaluation.

2.3. Model performance evaluation

We used three measures to evaluate the performance of our

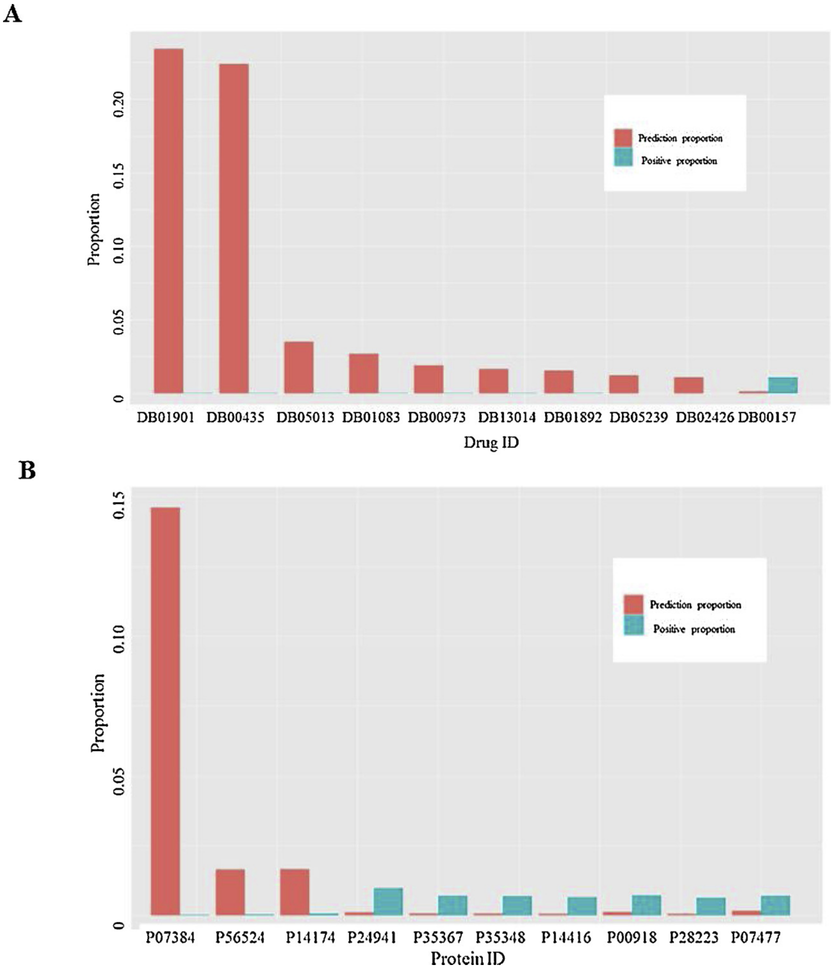


Fig. 9. Top 10 drugs and proteins with higher frequency of the interactions with other proteins and drugs respectively. **A:** DB01901 and DB00435 showed significantly higher proportions of the interactions with the proteins in the predicted drug-target interaction network. **B:** P07383 showed significantly higher proportions of the interactions with the drugs in the predicted drug-target interaction network.

classifiers. The first one is overall accuracy, which is the percentage of correctly predicted DTIs. The second one is the Receiver Operating Characteristics (ROC) curve, which illustrates the pattern of sensitivity (1-FNR (False negative rate)) and specificity (1-FPR (False positive rate)) at different cut-offs of the probability to predict a DTI to be a true interaction pair. Area under the curve (AUC) was calculated based on the ROC curve for each model to describe the quality of the classifiers. Lastly, the precision - recall curve (PRC) was calculated as the fraction of true positives among the full list of positive predictions, which provides a more accurate visual interpretability for imbalanced datasets (Saito and Rehmsmeier, 2015). We compared the PRC curves for each model.

2.4. Drug repurposing for breast cancer using the predicted drug-target interaction network

Breast cancer is the most common malignant cancer in women worldwide. It is known that family history is one of the major factor associated with breast cancer in women. However, the inherited components of breast cancer are not well established. Approximately 100 breast cancer risk loci have been identified in the GWAS to date. To translate these findings into a greater understanding of the mechanisms that influence disease risk, identification of the genes or non-coding RNAs that mediate these associations are required. Baxter et al. used Capture Hi-C (CHi-C) to annotate 63 of the established breast cancer risk loci (Baxter et al., 2018). They identified CHi-C interaction peaks involving 110 putative target genes mapping to 33 loci. Moreover, the breast cancer risk variants identified in the GWAS explain only a small fraction of the familial relative risk, and the genes responsible for these associations remain largely unknown. To identify novel risk loci and likely causal genes, Wu et al. (Wu et al., 2018) performed a transcriptome-wide association study to evaluate associations of genetically predicted gene expression with breast cancer risk in 122,977 cases and 105,974 controls of European ancestry. They used data from the Genotype-Tissue Expression Project (Lonsdale et al., 2013) to establish genetic models and predict gene expression in breast tissue. They evaluated the model performance using data from The Cancer Genome Atlas (Network, 2013). In total, 8597 genes were evaluated. They identified 179 genes associated with breast cancer risk.

In this study we combined the identified risk genes of Baxter's study (Baxter et al., 2018) and Wu's study (Wu et al., 2018), which included a total of 288 risk genes of breast cancer used for drug repurposing. We extracted the predicted DTIs with probability larger than or equal to 0.9995 from the predicted network that contained the breast cancer risk genes (proteins). We calculated the frequency of each drug interaction with the breast cancer risk genes based on these extracted DTIs. We repurposed the top drugs ranked by their frequency for breast cancer.

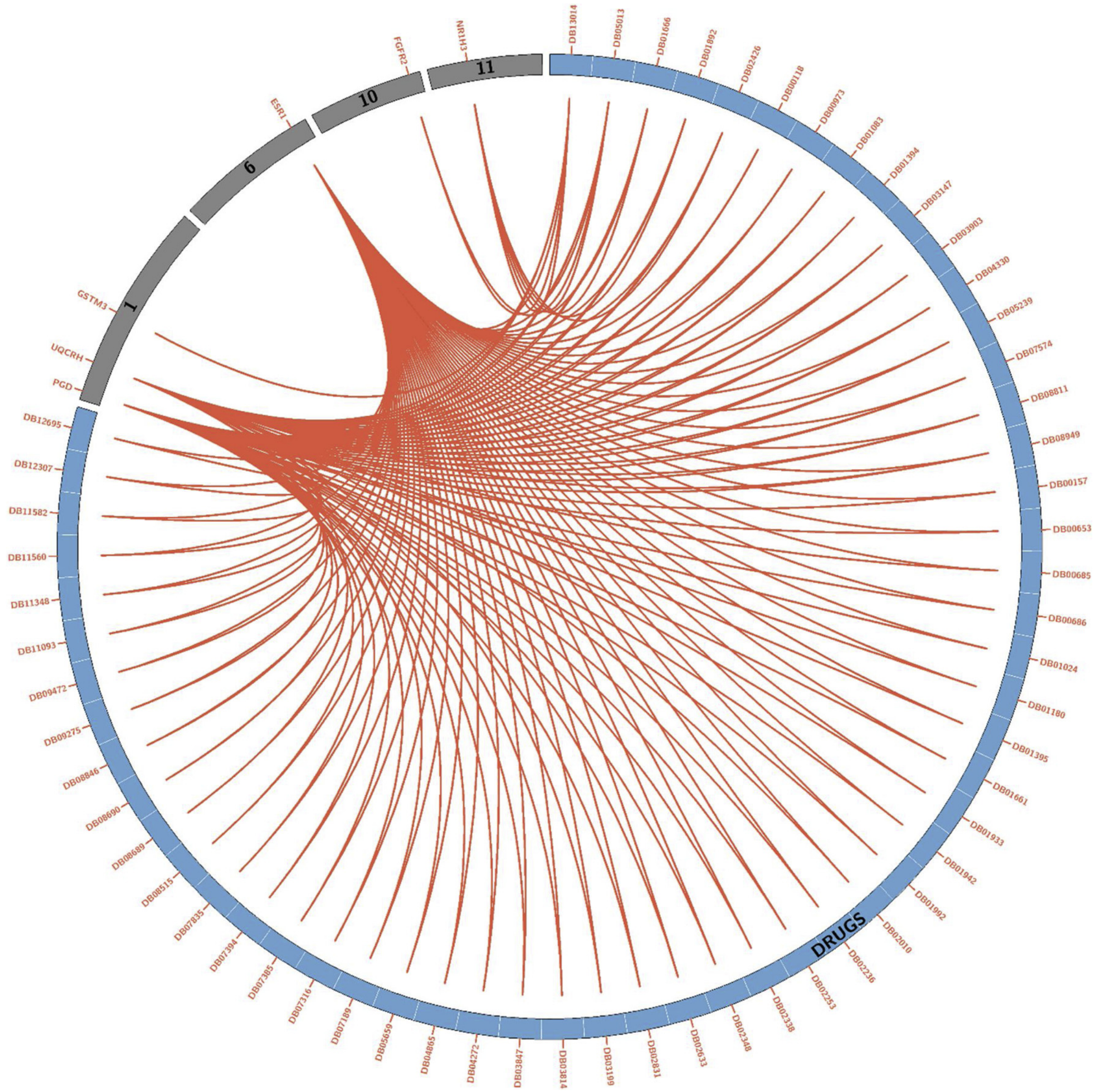


Fig. 10. 57 drugs interactions with 6 breast cancer risk genes in the predicted DTIs. Blue band locates drugs and gray band stands for the chromosomes.

3. Result

3.1. Standard logistic regression models

We fitted the SLG models using the top 100, 200, 500, 1000 and 2000 features selected by Wilcoxon rank sum test based on the training data. Table 1 shows the prediction accuracies and AUCs of our test set based on the trained models. Overall, models with more features had greater power to predict DTIs. The model had the best prediction performance when 2000 features were used in this study. We also tried models with more than 2000 features, but the models could not converge properly. In general, the models' performance decreased when the number of features used in the models decreased.

3.2. LASSO models

3.2.1. Parameter selection in the training data

Fig. 6**A–E** shows the AUCs of the five models on our training set based on 10-fold cross-validation. The largest AUCs were achieved for X-axis points called Lambda_min. Instead of using Lambda_min, we used Lambda_1se to fit the models. The Lambda_1se we selected was the largest value of Lambda such that the AUC was within 1 standard error (1se) of the maximum AUCs. Fig. 6**F** shows the results based on Lambda_1se for the five models. Although the Lambda_1se did not result in the best performance in the training set, it considered model uncertainty to make more reliable predictions for the test set.

3.2.2. Performance evaluation of the LASSO models

Given the Lambda_1se we selected for the five models, Table 2 shows the accuracies and AUCs of the five LASSO models in the training set based on 10-fold cross-validation and the test set. Fig. 7**A** and **B** shows the ROCs of the five models for the training set based on 10-fold cross-validation and the test set respectively. Generally speaking, the best model was Model 1 with all five feature sets. Model 2 and Model 3 with fewer numbers of features seemed weaker than other models. However, the model performances using protein TC information (Model 4) and protein domain information (Model 5) were remarkably improved. Better results were obtained from the test set than the training set, which indicated we avoided overfitting perhaps due to the model Lambda selection.

3.3. SVM models

We applied the SVM models using various parameters for C: 0.1,

Table 6
Top drugs repurposed for breast cancer.

Drug ID (Name)	Interacted genes	Diseases treated by the drugs	Drug-breast cancer association
DB13014 (Hypericin)	ESR1,UQCRH,GSTM3,FGFR2,PGD,NR1H3	Used for Human Immunodeficiency Virus (HIV) Infections and Cutaneous T-Cell Lymphoma (CTCL) (Wishart et al., 2006).	Hypericum perforatum has shown cytotoxic and apoptogenic effect on the MCF-7 human breast cancer cell line (Mirmalek et al., 2016).
DB05013 (Ingenol ebutate)	ESR1,UQCRH,FGFR2,PGD,NR1H3	Used for the topical treatment of actinic keratosis (Wishart et al., 2006).	Shown to be a potent anti-cancer drug and therapeutically effective in microgram quantities. Making it a promising candidate for use in breast cancer patients (Ogbourne et al., 2014).
DB01666 (D-Myo-Inositol-Hexasulphate)	ESR1,UQCRH,PGD,NR1H3	Myo-inositol hexasulfate is a potent inhibitor of Aspergillus ficuum phytase (Ullah and Sethumadhavan, 1998).	Experiment data of breast cancer cell lines treated in vitro with myo-Ins indicated that myo-Ins inhibits the principal molecular pathway supporting EMT in cancer cells (Bizzarri et al., 2016).
DB01892 (Hyperforin)	ESR1,UQCRH,PGD,NR1H3	Hyperforin is a possible antidepressant component (Chatterjee et al., 1998). It also induces hepatic drug metabolism through activation of the pregnane X receptor (Moore et al., 2000).	Hyperforin has shown anti-inflammatory and anti-carcinogenic properties (Koeberle et al., 2011).
DB02426 (Carboxyatractyloside)	ESR1,UQCRH,PGD,NR1H3	Carboxyatractyloside belongs to the class of organic compounds known as diterpene glycosides (Wishart et al., 2006).	Carboxyatractyloside is a specific inhibitors of ANT, and the expression of ANT isoforms is closely related to the energetic metabolic properties of tumoral cells (Chevrollier et al., 2011).

0.3, 0.5, 0.7 and 0.9. Table 3 shows the prediction accuracies and AUCs of our test set based on the trained models. Overall, all these models did not show much difference in prediction performance with various C values. Since smaller C values could tolerate less misclassification error from the training samples, C = 0.5 was considered to be the most suitable for an easier model to interpret.

3.4. Standard DNN model

We trained the standard deep neural network shown in Fig. 4 using a learning rate = 0.00001, dropout rate = 0.2 in the first two hidden layers and L2 regularization for 100,000 iterations, the accuracy and AUC on testing samples were 76% and 0.81 respectively, which were comparable to the results from SLG and LASSO models.

3.5. LASSO-DNN model

3.5.1. Feature extraction in the training data

To build the LASSO-DNN network, we first screened drug- and protein-specific features using the LASSO regression models on the training data, which were named Model 6 (M6) and Model 7 (M7), respectively. Table 4 shows the two models' accuracies, AUC and number of significant features based on Lambda_min and Lambda_1se. We observed that there were minor differences of the models' performance between Lambda_min and Lambda_1se. Hence, we used Lambda_min in both models for feature extraction since a larger number of significant features was obtained.

3.5.2. Performance evaluation of the LASSO-DNN model

For the 883 drug- and 1078 protein-specific features extracted from the LASSO models, we concatenated them as the input feature set for our new DNN model, called LASSO-DNN model (Fig. 5). We trained the network using learning rate = 0.00001, dropout rate = 0.2 in the first two hidden layers and L2 regularization, which achieved the AUC of 0.99 on the training set. The test set was first matched with the extracted features before feeding into the trained LASSO-DNN model. The accuracy and AUC of the trained LASSO-DNN model on the test data were 81% and 0.89, respectively. ROCs for each model were demonstrated in Fig. 8**A**. The precision - recall curve (PRC) illustrated the relationship between precision values and recall values. PRCs for each model were demonstrated in Fig. 8**B.** All these showed that the new DNN (LASSO-DNN) model achieved a better performance than those of our baseline LASSO, SLG, SVM and standard DNN models (Table 5**).**

3.6. Drug repurposing using the predicted drug-target interaction network

Using the trained LASSO-DNN model, we predicted all possible 18,138,422 DTIs combined from the 5134 drugs and the 3533 proteins. We built a drug-target interaction network based on the top 15,000 interactions from the predicted DTIs with a probability score cutoff 0.9995, which is a similar number of DTIs in Drugbank (14,792, see Fig. 2). The network included 3082 drugs and 3514 proteins. For the predicted and known drug-target interaction networks, we calculated the frequency of each drug interaction (Fig. 9A) and the frequency of each protein interaction (Fig. 9B). It is obvious that DB01901 and DB00435 were interacting proteins taking more than 20% of the interactions in the predicted network, and P07383 were interacting drugs taking more than 15% of the interactions in the predicted network. Hence, we excluded the interactions containing DB01901, DB00435 or P07383 in the predicted network for further exploration. The final predicted network included 5921 drug-target interactions with 1633 unique drugs and 608 unique proteins.

For the 288 breast cancer risk genes and the final predicted network, we extracted a subnetwork with 138 DTIs, including only 6 of the 288 breast cancer risk genes and 57 drugs. These 57 drugs had at least two predicted interactions with the 6 breast cancer risk genes as shown in Fig. 10. Table 6 shows the potentially repurposed drugs based on the subnetwork, which had interacted with at least 4 of the 6 breast cancer risk genes. Overall, all 5 of these drugs have shown evidence in relation to breast cancer treatment.

4. Conclusions

In this study we first applied SLG and LASSO regression models to predict DTIs. Our experiment showed that the LASSO-based regularized linear classification models achieved better results than the SLG models for predicting the DTIs. We also demonstrated that the LASSO models can work well for data sets with a very large number of features, which often result in overfitting in the SLG models. We explored the prediction power of different protein feature sets and showed that the best performance can be reached by integrating all feature sets. The standard DNN model was also applied using the full list of features. We found it had similar performance to the LASSO models. Finally, we introduced the LASSO-DNN model based on features extracted from the LASSO regression models fitted using the protein-specific and drug-specific features respectively. We showed that the LASSO-DNN model has better performance than the SLG, LASSO, SVM and the standard DNN models. We predicted all the combined drug-target pairs using the trained LASSO-DNN model. We repurposed potential drugs for breast cancer using the predicted drug-target interaction network.

One of the limitations in our study is that the modeling strategy is sensitive to the selection of negative DTIs since there are no existing gold standard negative DTIs available. To avoid selecting potential positive DTIs as our negative DTIs, we tried a similarity-based approach to filter out those potential positive DTIs from our negative DTIs. Our models may vary when different cut-offs are applied in selection of potential positive DTIs or a different method to be used for generating negative DTIs. Given this consideration, we will explore other machine learning approaches in selecting the negative DTIs in the future.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported in part by the Canadian Breast Cancer Foundation; the Natural Sciences and Engineering Research Council of Canada; the Manitoba Research Health Council and the University of Manitoba. The authors confirm that the funders had no influence over the study design, the content of the article, or selection of this journal.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., Brain, G., 2016. TensorFlow: a system for large-scale machine learning. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI' 16) 265–284. https://doi.org/10.1038/nn.3331.

Ashburn, T.T., Thor, K.B., 2004. Drug repositioning: identifying and developing new uses for existing drugs. Nat. Rev. Drug Discov. 3, 673–683. https://doi.org/10.1038/nrd1468.

Bajusz, D., Rácz, A., Héberger, K., 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J. Chem. Inform. 7 (1), 20. https://doi.org/10.1186/s13321-015-0069-3.

Baxter, J.S., Leavy, O.C., Dryden, N.H., Maguire, S., Johnson, N., Fedele, V., Simigdala, N., Martin, L.A., Andrews, S., Wingett, S.W., Assiotis, I., Fenwick, K., Chauhan, R., Rust, A.G., Orr, N., Dudbridge, F., Haider, S., Fletcher, O., 2018. Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. Nat. Commun. 9 (1), 2018. https://doi.org/10.1038/s41467-018-03411-9.

Bizzarri, M., Dinicola, S., Bevilacqua, A., Cucina, A., 2016. Broad Spectrum anticancer activity of myo-inositol and inositol hexakisphosphate. Int. J. Endocrinol. https://doi.org/10.1155/2016/5616807. 27795708.

Chatterjee, S.S., Bhattacharya, S.K., Wonnemann, M., Singer, A., Müller, W.E., 1998. Hyperforin as a possible antidepressant component of hypericum extracts. Life Sci. 63 (6), 499–510. https://doi.org/10.1016/S0024-3205(98)00299-9.

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W., Huang, J., Tang, Y., 2012. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput. Biol. 8 (5), e1002503. https://doi.org/10.1371/journal.pcbi.1002503.

Chevrollier, A., Loiseau, D., Reynier, P., Stepien, G., 2011. Adenine nucleotide translocase 2 is a key mitochondrial protein in cancer metabolism. Biochim. Biophys. Acta Bioenerg. 1807 (6), 562–567. https://doi.org/10.1016/j.bbabio.2010.10.008.

Ezzat, A., Wu, M., Li, X.-L., Kwoh, C.-K., 2017. Drug-target interaction prediction using ensemble learning and dimensionality reduction. Methods 129, 81–88. https://doi.org/10.1016/j.ymeth.2017.05.016.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1–22. https://doi.org/10.18637/jss.v033.i01.

Gilbert, J., Henske, P., Singh, A., 2003. Rebuilding big pharma's business model. Vivo, Bus. Med. Rep. 21, 73–80.

Hamanaka, M., Taneishi, K., Iwata, H., Ye, J., Pei, J., Hou, J., Okuno, Y., 2017. CGBVS-DNN: prediction of compound-protein interactions based on deep learning. Mol. Inform. 36 (1-2), 1600045. https://doi.org/10.1002/minf.201600045.

He, Z., Zhang, J., Shi, X., Hu, L., Kong, X., Cai, Y., Chou, K., 2010. Predicting drug-target interaction networks based on functional groups and biological features. PLoS One 5, e9603. https://doi.org/10.1371/journal.pone.0009603.

Health, R. on T.G.-B.R. for, Policy, B. on H.S, Medicine, I. of, 2014. Drug Repurposing and Repositioning: Workshop Summary, the National Academies Collection: Reports Funded by National Institutes of Health.2015. pp. 8. https://doi.org/10.17226/18731.

Horrobin, D.F., 2001. Realism in drug discovery - could cassandra be right? Nat. Biotechnol. 19, 1099–1100. https://doi.org/10.1038/nbt1201-1099.

Kingma, D.P., Ba, J.L., 2015. Adam: a method for stochastic optimization. Int. Conf. Learn. Represent. 2015, 1–15. https://doi.org/10.1145/1830483.1830503.

Koeberle, A., Rossi, A., Bauer, J., Dehm, F., Verotta, L., Northoff, H., Sautebin, L., Werz, O., 2011. Hyperforin, an anti-inflammatory constituent from St. john's wort, inhibits microsomal prostaglandin E2 Synthase-1 and suppresses prostaglandin E2 formation in vivo. Front. Pharmacol. 2, 7. https://doi.org/10.3389/fphar.2011.00007.

Leung, M.K.K., Delong, A., Alipanahi, B., Frey, B.J., 2016. Machine learning in genomic medicine: a review of computational problems and data sets. Proc. IEEE 104, 176–197. https://doi.org/10.1109/JPROC.2015.2494198.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, Daniel, Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N.J., Nicolae, D.L., Gamazon, E.R., Im, H.K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T., Wen, X., Dermitzakis, E.T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M., Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., Sturcke, A., Paschal, J., Anderson, J.M., Wilder, E.L., Derr, L.K., Green, E.D., Struewing, J.P., Temple, G., Volpi, S., Boyer, J.T., Thomson, E.J., Guyer, M.S., Ng, C., Abdallah, A., Colantuoni, D., Insel, T.R., Koester, S.E., Roger Little, A., Bender, P.K., Lehner, T., Yao, Y., Compton, C.C., Vaught, J.B., Sawyer, S., Lockhart, N.C., Demchok, J., Moore, H.F., 2013. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 45 (6), 580. https://doi.org/10.1038/ng.2653.

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., Peng, J., Chen, L., Zeng, J., 2017. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat. Commun. 8 (1), 573. https://doi.org/10.1038/s41467-017-00680-8.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., MayPendlington, Z., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., Parkinson, H., 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 45, D896–D901. https://doi.org/10.1093/nar/gkw1133.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2014. Misc Functions of the Department of Statistics (e1071). version 1.6-2. TU Wien. R Packag. https://doi.org/citeulike-article-id:9958545.

Min, S., Lee, B., Yoon, S., 2017. Deep learning in bioinformatics. Brief. Bioinform. 18, 851–869. https://doi.org/10.1093/bib/bbw068.

Mirmalek, S.A., Azizi, M.A., Jangholi, E., Yadollah-Damavandi, S., Javidi, M.A., Parsa, Y., Parsa, T., Salimi-Tabatabaee, S.A., Ghasemzadeh kolagar, H., Alizadeh-Navaei, R., 2016. Cytotoxic and apoptogenic effect of hypericin, the bioactive component of Hypericum perforatum on the MCF-7 human breast cancer cell line. Cancer Cell Int. 16 (1), 3. https://doi.org/10.1186/s12935-016-0279-4.

Moore, L.B., Goodwin, B., Jones, Sa, Wisely, G.B., Serabjit-Singh, C.J., Willson, T.M., Collins, J.L., Kliewer, Sa, 2000. St. John's wort induces hepatic drug metabolism through activation of the pregnane X receptor. Proc. Natl. Acad. Sci. U. S. A. 97, 7500–7502. https://doi.org/10.1073/pnas.130155097.

Napolitano, F., Zhao, Y., Moreira, V.M., Tagliaferri, R., Kere, J., D'Amato, M., Greco, D., 2013. Drug repositioning: a machine-learning approach through data integration. J. Cheminform. 5 (1), 30. https://doi.org/10.1186/1758-2946-5-30.

Network, T.C.G.A.R., 2013. The Cancer Genome Atlas pan-cancer analysis project. Nat. Genet. 45, 1113–1120. https://doi.org/10.1038/ng.2764.

Ogbourne, S. M. U.S. Patent No. 8,653,133. Washington, DC: U.S. Patent and Trademark Office. (2014).

Pritchard, J.L.E., O'Mara, T.A., Glubb, D.M., 2017. Enhancing the promise of drug repositioning through genetics. Front. Pharmacol. 8, 896. https://doi.org/10.3389/fphar.2017.00896.

Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10, e0118432. https://doi.org/10.1371/journal.pone.0118432.

Sanseau, P., Agarwal, P., Barnes, M.R., Pastinen, T., Richards, J.B., Cardon, L.R., Mooser, V., 2012. Use of genome-wide association studies for drug repositioning. Nat. Biotechnol. 30 (4), 317. https://doi.org/10.1038/nbt.2151.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from Overfitting. J. Mach. Learn. Res. 15, 1929–1958. https://doi.org/10.1214/12-AOS1000.

Sushko, I., Novotarskyi, S., Körner, R., Pandey, A.K., Rupp, M., Teetz, W., Brandmaier, S., Abdelaziz, A., Prokopenko, V.V., Tanchuk, V.Y., Todeschini, R., Varnek, A., Marcou, G., Ertl, P., Potemkin, V., Grishina, M., Gasteiger, J., Schwab, C., Baskin, I.I., Palyulin, V.A., Radchenko, E.V., Welsh, W.J., Kholodovych, V., Chekmarev, D., Cherkasov, A., Aires-De-Sousa, J., Zhang, Q.Y., Bender, A., Nigsch, F., Patiny, L., Williams, A., Tkachenko, V., Tetko, I.V., 2011. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J. Comput. Aided Mol. Des. 25, 533–554. https://doi.org/10.1007/

s10822-011-9440-2.

Ullah, A.H.J., Sethumadhavan, K., 1998. Myo-inositol hexasulfate is a potent inhibitor of Aspergillus ficuum Phytase. Biochem. Biophys. Res. Commun. 251, 260–263. https://doi.org/10.1006/bbrc.1998.9456.

Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., Lu, H., 2017. Deep- learning-Based drug-target interaction prediction. J. Proteome Res. 16, 1401–1409. https://doi.org/10.1021/acs.jproteome.6b00618.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L.Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Miller, V., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L., Yaschenko, E., 2007. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 35 (suppl_1), D5–D12. https://doi.org/10.1093/nar/gkl1031.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., MacIejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, Di., Pon, A., Knox, C., Wilson, M., 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 46, D1074–D1082. https://doi.org/10.1093/nar/gkx1037.

Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., Bolla, M.K., Shu, X.-O., Lu, Y., Cai, Q., Al-Ejeh, F., Rozali, E., Wang, Q., Dennis, J., Li, B., Zeng, C., Feng, H., Gusev, A., Barfield, R.T., Andrulis, I.L., Anton-Culver, H., Arndt, V., Aronson, K.J., Auer, P.L., Barrdahl, M., Baynes, C., Beckmann, M.W., Benitez, J., Bermisheva, M., Blomqvist, C., Bogdanova, N.V., Bojesen, S.E., Brauch, H., Brenner, H., Brinton, L., Broberg, P., Brucker, S.Y., Burwinkel, B., Caldés, T., Canzian, F., Carter, B.D., Castelao, J.E., Chang-Claude, J., Chen, X., Cheng, T.-Y.D., Christiansen, H., Clarke, C.L., Collée, M., Cornelissen, S., Couch, F.J., Cox, D., Cox, A., Cross, S.S., Cunningham, J.M., Czene, K., Daly, M.B., Devilee, P., Doheny, K.F., Dörk, T., dos-Santos-Silva, I., Dumont, M., Dwek, M., Eccles, D.M., Eilber, U., Eliassen, A.H., Engel, C., Eriksson, M., Fachal, L., Fasching, P.A., Figueroa, J., Flesch-Janys, D., Fletcher, O., Flyger, H., Fritschi, L., Gabrielson, M., Gago-Dominguez, M., Gapstur, S.M., García-Closas, M., Gaudet, M.M., Ghoussaini, M., Giles, G.G., Goldberg, M.S., Goldgar, D.E., González-Neira, A., Guénel, P., Hahnen, E., Haiman, C.A., Håkansson, N., Hall, P., Hallberg, E., Hamann, U., Harrington, P., Hein, A., Hicks, B., Hillemanns, P., Hollestelle, A., Hoover, R.N., Hopper, J.L., Huang, G., Humphreys, K., Hunter, D.J., Jakubowska, A., Janni, W., John, E.M., Johnson, N., Jones, K., Jones, M.E., Jung, A., Kaaks, R., Kerin, M.J., Khusnutdinova, E., Kosma, V.-M., Kristensen, V.N., Lambrechts, D., Le Marchand, L., Li, J., Lindström, S., Lissowska, J., Lo, W.-Y., Loibl, S., Lubinski, J., Luccarini, C., Lux, M.P., MacInnis, R.J., Maishman, T., Kostovska, I.M., Mannermaa, A., Manson, J.E., Margolin, S., Mavroudis, D., Meijers-Heijboer, H., Meindl, A., Menon, U., Meyer, J., Mulligan, A.M., Neuhausen, S.L., Nevanlinna, H., Neven, P., Nielsen, S.F., Nordestgaard, B.G., Olopade, O.I., Olson, J.E., Olsson, H., Peterlongo, P., Peto, J., Plaseska-Karanfilska, D., Prentice, R., Presneau, N., Pylkäs, K., Rack, B., Radice, P., Rahman, N., Rennert, G., Rennert, H.S., Rhenius, V., Romero, A., Romm, J., Rudolph, A., Saloustros, E., Sandler, D.P., Sawyer, E.J., Schmidt, M.K., Schmutzler, R.K., Schneeweiss, A., Scott, R.J., Scott, C.G., Seal, S., Shah, M., Shrubsole, M.J., Smeets, A., Southey, M.C., Spinelli, J.J., Stone, J., Surowy, H., Swerdlow, A.J., Tamimi, R.M., Tapper, W., Taylor, J.A., Terry, M.B., Tessier, D.C., Thomas, A., Thöne, K., Tollenaar, R.A.E.M., Torres, D., Truong, T., Untch, M., Vachon, C., Van Den Berg, D., Vincent, D., Waisfisz, Q., Weinberg, C.R., Wendt, C., Whittemore, A.S., Wildiers, H., Willett, W.C., Winqvist, R., Wolk, A., Xia, L., Yang, X.R., Ziogas, A., Ziv, E., Dunning, A.M., Pharoah, P.D.P., Simard, J., Milne, R.L., Edwards, S.L., Kraft, P., Easton, D.F., Chenevix-Trench, G., Zheng, W., 2018. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nat. Genet. 50, 1. https://doi.org/10.1038/s41588-018-0132-x.

Xiao, N., Cao, D.S., Zhu, M.F., Xu, Q.S., 2015. Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics. 1857–1859. https://doi.org/10.1093/bioinformatics/btv042.

Yamanishi, Y., Kotera, M., Moriya, Y., Sawada, R., Kanehisa, M., Goto, S., 2014. DINIES: drug-target interaction network inference engine based on supervised analysis. Nucleic Acids Res. 42 (W1), W39–W45. https://doi.org/10.1093/nar/gku337.

Zhao, K., So, H.-C., 2017. A Machine Learning Approach to Drug Repositioning Based on Drug Expression Profiles: Applications to Schizophrenia and depression/anxiety Disorders. arXiv preprint arXiv,1706.03014. .